Proposal for a Symposium at the Spring 2011 SREE Conference

Symposium Title:
Statistical Approaches to Studying Mediator Effects In Education Research

Chair Person:
Howard Bloom (MDRC)  Howard.Bloom2@mdrc.org

Choice of Conference Section:
Research Methods

Proposed Papers (in presentation order):
- “Under What Assumptions do Site-by-Treatment Instruments Identify Average Causal Effects?” by Reardon and Raudenbush
- “Assessing Compliance-Effect Bias in the Two Stage Least Squares Estimator” by Reardon, Unlu, Zhu, and Bloom
- “Mediation and Spillover Effects in Group-Randomized Trials with Application to the 4Rs Evaluation” by VanderWeele, Hong, Jones and Brown

Discussant:
Peter Schochet (Mathematica Policy Research)
Symposium Justification:

The growing use of rigorous methods to estimate effects of educational interventions is now making it possible to accumulate strong evidence about the effects of such efforts. However this progress does not solve the tougher problem of determining the effects of the intermediate causal factors (“mediators”) that comprise interventions’ “active ingredients.” Given the importance of taking this next step, our proposed panel will examine alternative methods for identifying and estimating causal effects of mediators. The two approaches considered are: (1) instrumental variables analysis of multi-site randomized trials and (2) multi-level mediational analyses in the presence of spillover effects.

Two papers focus on the first approach. “Under What Conditions do Site-by-Treatment Interactions Identify Average Causal Effects,” by Reardon and Raudenbush, articulates the assumptions that must be met in order for multiple instrumental variables produced by site-by-treatment interactions to provide valid estimates of mediator effects. This approach is growing in popularity but its underlying assumptions are not well or widely understood, particularly in situations where the effect of a mediator varies across sites or individuals. By theoretically exploring this issue, the first paper makes a major contribution to the methodological literature.

The second paper by Reardon, Unlu, Zhu and Bloom, “Assessing Compliance-Effect Bias in the Two Stage Least Squares Estimator,” builds directly on the first by quantifying the bias that occurs when a central assumption of multi-site IV analysis is violated. This assumption requires that there be no correlation across sites or individuals between the effect of treatment status on the mediator and the effect of the mediator on the outcome. For example, if an intervention were intended to increase the number of minutes per day that teachers use a particular kind of instruction and the outcome were students’ achievement, the required assumption would be that variation across sites (or schools or classrooms) in the effect of the intervention on the amount of this instruction and variation across sites (or schools or classrooms) in the effect of this instruction on student achievement be independent. The paper presents closed-form expressions for the bias that results when this correlation is nonzero plus simulations that explore how this “compliance effect bias” varies as a function of key parameters.

The third paper, “Mediation and Spillover Effects in Group-Randomized Trials with Application to the 4Rs Evaluation,” by VanderWeele, Hong, Jones and Brown, presents a multi-level modeling approach to studying mediators in situations where spillovers exist among sample members. This situation occurs when response to an intervention by some sample members affect the response (and thus outcomes) of other sample members. The authors present a multi-level model that account for the analytic problem when spillovers exist. This model can provide valid estimates of the effect of a given mediator and of the spillover effects themselves. The approach thus has potential substantive and methodological benefits. However it requires stringent assumptions which the authors articulate. Use of the approach is illustrated by its application to a school-randomized trial of an intervention to improve socio-emotional outcomes for young students (4Rs).
Title: Under What Assumptions do Site-by-Treatment Instruments Identify Average Causal Effects?

Author(s): Sean F. Reardon (Stanford University, contact author) 
Stephen W. Raudenbush (University of Chicago)
Abstract Body

Background / Context:

In canonical applications of the instrumental variable method, exogenously determined exposure to an instrument (such as random assignment to a treatment condition) induces exposure to a mediating process that in turn causes a change in a later outcome. A crucial assumption known as the exclusion restriction is that the hypothesized instrument can influence the outcome only through its influence on exposure to the mediator of interest (Heckman & Robb, 1985; Imbens & Angrist, 1994). It may be the case, however, that multiple mediators operate jointly to influence the outcome, in which case a single instrument will not suffice to identify the causal effects of interest.

To cope with this problem, analysts have recently exploited the fact that a causal process is often replicated across multiple sites, generating the possibility of multiple instruments in the form of site-by-instrument interactions. These multiple instruments can, in principle, enable the investigator to identify the impact of multiple programs or treatments regarded as the mediators of the effect of an instrument. Kling, Liebman, and Katz (2007), for example, used random assignment in the Moving to Opportunity (“MTO”) study as an instrument to estimate the impact of neighborhood poverty on health, social behavior, education, and economic self-sufficiency of adolescents and adults. Reasoning that the instrument might affect outcomes through mechanisms other than neighborhood poverty, they control for a second mediator, use of the randomized treatment voucher. To do so, they capitalize on the replication of the MTO experiment in five cities, generating ten instruments (the five sites generate ten site-by-treatment interactions as instruments because there were three randomly assigned treatment conditions per site) to identify the impact of the two mediators of interest, neighborhood poverty and experimental compliance. Using a similar strategy, Morris, Duncan, and Rodriguez (2010) used data from 16 implementations of welfare-to-work experiments to identify the impact of family income, average hours worked, and receipt of welfare as mediators.

Clearly, this strategy for generating multiple instruments has potentially great appeal in research on causal effects in social science. For example, Spybrook (2009) found that, among 75 large-scale experiments funded by the US Institute of Education Sciences over the past decade, the majority were multi-site studies in which randomization occurred within sites. In principle, these data could yield a wealth of new knowledge about causal effects in education policy. It is essential, however, that researchers understand the assumptions required to pursue this strategy successfully. To date, we know of no complete account of these assumptions.

Purpose / Objective / Research Question / Focus of Study:

Our purpose is to clarify the assumptions that must be met if this “multiple site, multiple mediator” strategy, hereafter referred to as “MSMM,” is to identify the average causal effects (ATE) in the populations of interest.
Setting:
N/A

Population / Participants / Subjects:
N/A

Intervention / Program / Practice:
N/A

Significance / Novelty of study:

We are aware of only two studies that rely on site-by-treatment interactions to generate multiple instruments to estimate the impacts of multiple potential mediators (Kling et al., 2007; Morris et al., 2010). Neither of these studies discusses the assumptions underlying the models they fit. Given the large number of multi-site randomized trials that have been conducted in education and in other fields (Spybrook, 2009), such models are likely to become increasingly appealing as a means to estimating the effects of multiple hard-to-randomize potential mechanisms. Ours is the first paper that systematically describes the identifying assumptions of such models.

Statistical, Measurement, or Econometric Model:

We begin by delineating the assumptions required for identification in the case of a single instrument and a single mediator within a single-site study. Unlike Angrist, Imbens, and Rubin (1996) (hereafter AIR), we consider the general case where the mediator may be continuous or multi-valued. In this general case, the assumptions required for identification of the average treatment effect differ somewhat from those AIR (1996) describe for the binary mediator case. We link our discussion to recent papers describing the correlated random coefficient model, and show that the CRC model is identified by instrumental variables using a weaker assumption than that described by Heckman and Vytlacil (1998) and Wooldridge (2003).

Following a discussion of the single, site, single mediator case, we then consider the case of multiple sites with a single mediator before turning to the case of primary interest: MSMM. Finally, we generalize these results to any setting in which multiple instruments identify the impact of multiple mediators.

Usefulness / Applicability of Method:

Researchers often want to understand the mechanisms through which a specific treatment, program, or policy operates. Although it may be feasible to randomize individuals to specific treatments or programs, it is often not feasible to assign individuals to processes that are hypothesized to be mediators. For example, in the Reading First Impact Study (Gamse et al.,
2008), a regression discontinuity design provided exogenous variation in Reading First study, enabling the researchers to estimate the effects of the Reading First program as a whole. However, the program was hypothesized to work through its impacts on five distinct dimensions of teachers’ reading instruction practices. These practices could not be randomly assigned, however (because the researchers cannot control what teachers do in the classroom). One could imagine using the site-by-instrument interactions in an IV model to estimate the effects of specific instructional practices. There are many such cases in educational and social science research. Thus, a clearer understanding of the necessary assumptions (and the consequences of their failure) may improve the quality of research on hard-to-randomize educational processes and mechanisms.

Research Design:

We consider the case of a study design in which there are multiple sites, indexed by $s \in \{1, 2, ..., J\}$. In each site $s$, individuals are assigned to one of multiple possible values of a treatment, $T$, where $T$ may be binary or continuous. The treatment $T$ may affect a vector of $P$ mediators, $M = \{M_1, M_2, ..., M_P\}$; each of the mediators may also be binary or continuous. We describe the person-specific effect of $T$ on a mediator $M_p$ as the person-specific “compliance” with respect to mediator $M_p$. Each of the mediators may each affect an outcome, $Y$. We are interested in estimating the average effect of each of the mediators on the outcome $Y$.

Data Collection and Analysis:

N/A

Findings / Results:

We identify 9 assumptions that must be met in order that the MSMM IV model identifies the average effects of each mediator in the population of interest. These are:

1. **Stable unit treatment value assumption** (SUTVA): each unit has one and only one potential outcome under each treatment condition: that is, for a population of size $n$, $Y_i(t_1, t_2, ..., t_n) = Y_i(t_i)$ for all $i \in \{1, 2, ..., n\}$. In the IV model, this standard SUTVA assumption is actually composed of two distinct SUTVA assumptions:
   a. For each mediator $p$, each unit $i$ has one and only one potential value of $M_{pi}$ for each treatment condition $t$: that is, for a population of size $n$, $M_{pi}(t_1, t_2, ..., t_n) = M_{pi}(t_i)$ for all $i \in \{1, 2, ..., n\}$.
   b. Each unit $i$ has one and only one potential outcome value of $Y_i$ for each vector of mediators $M_i = \{M_{1i}, M_{2i}, ..., M_{pi}\}$: that is, for a population of size $n$, $Y_i(M_1, M_2, ..., M_n) = Y_i(M_i)$ for all $i \in \{1, 2, ..., n\}$.
2. **Exclusion restriction**: the treatment $T$ affects $Y$ only through its impact on the set of $P$ mediators, $\mathbf{M} = \{M_1, M_2, ..., M_P\}$. That is, $Y(t) = Y(t, \mathbf{M}(t)) = Y(\mathbf{M}(t))$.

3. **Person-specific linearity of each mediator in $T$**: the person-specific effect of $T$ on each mediator $M_p$ is linear. That is, $M_p(t) = M_p(0) + t \Gamma_p$ for each $p$.

4. **Person-specific linearity in $\mathbf{M}$**: the person-specific effect of each mediator $M_p$ is linear. That is, $Y(\mathbf{M}(t)) = Y(\mathbf{M}(0)) + \sum_{p=1}^{P} \Delta_p(t \Gamma_p)$.

5. **Ignorable within-site treatment assignment**: the assignment of the instrument (the treatment, in our notation) must be independent of the potential outcomes within each site: $T \perp Y(t)|s, T \perp \mathbf{M}(t)|s, \forall t, s$.

6. **No average within-site compliance-effect covariance**: $E\left(Cov_s(\Delta_p, \Gamma_p)\right) = 0, \forall p \in \{1, 2, ..., P\}$. A simpler, but stronger, assumption is that there is no within-site compliance-effect covariance in any site: $Cov_s(\Delta_p, \Gamma_p) = 0, \forall p \in \{1, 2, ..., P\}, \forall s \in \{1, 2, ..., S\}$.

7. **Site-by-mediator compliance matrix has sufficient rank**: In particular, if $G$ is the $S \times P$ matrix of the $gamma_{sp}$'s, then $\text{rank}(G) = P$. This implies three specific conditions:
   a. The compliance of at least $P - 1$ of the mediators varies across sites. That is, $Var\left(y_{ps}\right) = 0$, for at least one $p \in \{1, 2, ..., P\}$.
   b. There are at least as many sites as mediators: $P \leq S$.
   c. There is some subset of $Q$ site-specific average compliance vectors, $\Gamma_s = \{\gamma_{1s}, \gamma_{2s}, ..., \gamma_{ps}\}$, where $S \geq Q \geq P$, that are linearly independent.

8. **Parallel mediators**: assignment to $T$ does not influence a given mediator $M_p$ through any other mediator $M_q, q \in \{1, 2, ..., P\}$. That is, the mediators do not influence one another. That is, $M_p(t, M_1, ..., M_p-1, M_{p+1}, ..., M_P) = M_p(t), \forall p \in \{1, 2, ..., P\}$.

9. **Mean independence of the site-average compliances and effects**: Within each site $s$, let $y_{ps} = E[\Gamma_p|s]$ and $\delta_{ps} = E[\Delta_p|s]$ be the site-average compliance and effect of mediator $p$, respectively. Likewise, let $\gamma_p = E[\Gamma_p]$ and $\delta_p = E[\Delta_p]$ be the average compliance and average effect of mediator $p$ in the population. Then we assume that the site-average effects are independent of the site average compliances. That is, $E[\delta_{ps}|\gamma_{1s}, \gamma_{2s}, ..., \gamma_{ps}] = E[\delta_{ps}] = \delta_p \forall p \in \{1, ..., P\}$.

Some of these are familiar assumptions from the standard IV model in the case where there is a single site ($J = 1$) and a single mediator ($P = 1$). In this case, assumptions 1-7 are sufficient for the IV model to identify the average effect of the mediator on $Y$ in the population. Assumptions 8 and 9 pertain only when there are multiple mediators (assumption 8) or multiple sites (assumption 9). Moreover, in this case, if both $T$ and $M$ are binary, then assumptions 3 and
4 are unnecessary; assumptions 1, 2, 5, 6 and 7 are sufficient to identify the average effect of $M$ on $Y$ in the population.

We note that assumptions 1, 2, 5, and 7 are generalizations of four of the five IV assumptions identified by Angrist, Imbens, and Rubin (1996) [note that in the case when $S = P = 1$, the sufficient rank assumption is equivalent to the “nonzero average causal effect of $T$ on $M$” assumption made by AIR]. Their fifth assumption, montonicity, is notably absent from our set of assumptions, however. We show that the monotonicity assumption is relevant only in the case where both $T$ and $M$ are binary and where we wish to identify the average effect of $M$ on $Y$ among the population of “compliers”, in Angrist, Imbens, and Rubins’ (1996) terminology. If we wish to identify the average effect of $M$ on $Y$ in the population (arguably an estimand of more general interest), then no monotonicity assumption is needed. Instead, we require assumption 6, the no compliance-effect covariance assumption. Moreover, this assumption is relevant even when $T$ and $M$ are not binary. Thus, our set of assumptions is more generally applicable than the AIR assumptions, and identifies an estimand that is of more general interest.

We note also that assumption 6 (no within-site compliance-effect covariance) is weaker than the assumption that Heckman and Vytlacil (1998) and Wooldridge (2003) make in order to identify average effects in the correlated random coefficients model. We show that the assumption of no compliance effect covariance (along with the other standard IV assumptions) is sufficient to identify the average effects using an IV model when the data are generated from a correlated random effects process.

Conclusions:

Our investigation of the assumptions of the multiple-mediator, multiple-site IV model demonstrates that such models rely on a large number of non-trivial assumptions. Of most importance are the assumptions regarding the relationship between compliance and effect. Any correlation—whether within or between sites—between compliance (with respect to any mediator) and effect (again, of any mediator) will potentially bias the estimates of the effect of any mediator. Because individuals or groups (sites) with the most to gain from a specific practice (mediator) may be more likely to comply with treatment assignment, there is good reason to worry that compliance-effect covariances are non-zero in many cases.
Appendices

Not included in page count.

Appendix A. References


Title: Assessing Compliance-Effect Bias in the Two Stage Least Squares Estimator

Author(s): Sean Reardon (Stanford University) sreardon@suse.stanford.edu
Fatih Unlu (Abt Associates, Inc., contact author) Fatih_Unlu@abtassoc.com
Pei Zhu (MDRC) Pei.Zhu@mdrc.org
Howard Bloom (MDRC) Howard.Bloom2@mdrc.org
Abstract Body

Background / Context:

In the past eight years, education research has taken a quantum leap forward based on a large and growing number of high-quality randomized field trials and regression discontinuity studies of the effects of educational interventions. Most of this new research and existing methodologies for conducting it focus on the response of student academic outcomes to specific educational interventions. Such information is invaluable and can provide a solid foundation for accumulating much-needed knowledge. However, this information only indicates how well specific interventions (which comprise complex bundles of features) work for specific students in specific settings. Therefore by itself, the information is not sufficient to ascertain “What works best for whom, when and why?” And it is this more comprehensive understanding of educational interventions that is needed to guide future policy and practice.

In other words, it is necessary to “unpack the black boxes” being tested by randomized experiments or high-quality quasi-experiments in order to learn how best to improve the education—and thus life chances—of students in the U.S., especially those who are economically disadvantaged. This unpacking job comprises learning more about the relative effectiveness of the active ingredients of educational interventions (their mediators) and learning more about factors that influence the effectiveness of these interventions (their moderators).

Now that multi-site randomized experiments and rigorous quasi-experiments have been shown to be feasible for educational research it is an opportune time to begin to explore these subtler and more complex questions.

Of particular relevance for the present paper is the use of instrumental variables analysis in the context of multi-site randomized experiments or quasi-experiments to study the effects of mediating variables on final outcomes. In particular, recent applications of the approach have begun to use it to explore causal effects of one or more mediating factors. For example, data from a randomized trial of subsidies for public housing residents to stimulate movement to lower-poverty neighborhoods were used to study the effects of neighborhood poverty on child outcomes (Kling, Liebman, and Katz, 2007). Using a similar strategy, Morris, Duncan, and Rodrigues (2010) used data from 16 implementations of welfare-to-work experiments to identify the impact of family income, average hours worked, and receipt of welfare as mediators.

Even though this strategy for generating multiple instruments has potentially great appeal in research on causal effects through multiple mediators in education policy, the conditions under which this strategy can be used to identify the average treatment effect (ATE) has not been addressed until recently. Specifically, Reardon and Raudenbush (2010) fills this vacuum by

---

2 Spybrook and Raudenbush (2009) identified 75 randomized studies of a broad range of interventions and Gamse et al. (2008) and Jackson et al. (2007) report on regression discontinuity studies of the federal Reading First and Early Reading First programs.

3 An important exception involves a series of randomized tests of interventions for improving students’ social and emotional outcomes (Jones, Brown, and Aber, 2008 and Haegerick and Metz, under review).

4 Cook (2001) speculates about why, until recently, the education research community strenuously resisted randomized experiments.

demonstrating that a number of assumptions above and beyond the canonical instrumental variable analysis assumptions (Angrist, Imbens, and Rubin, 1996) are needed to identify the average treatment effect in the case of a multi-site study in which an instrument may affect the outcome through multiple mediators. One of their key assumptions is that the effect of the treatment (instrument) on a mediator (“compliance”) should not be correlated with the effect of that mediator on the outcome of interest (“effect”), i.e., no compliance-effect covariance. The proposed paper zooms in on this assumption and assesses the properties of the most common instrumental variable estimator (the two-stage least squares, or 2SLS, estimator) when there is compliance-effect covariance.

**Purpose / Objective / Research Question / Focus of Study:**

The proposed paper studies the bias in the 2SLS estimator that is caused by the compliance-effect covariance (hereafter, the compliance-effect bias). It starts by deriving the formula for the bias in an infinite sample (i.e., in the absence of finite sample bias) under different circumstances. Specifically, it considers the following cases:

a) A single site study with one mediator;

b) A multiple site study with one mediator; and

c) A multiple site study with multiple mediators.

The formulas demonstrate how the magnitude of the compliance-effect bias varies with different parameters (e.g., compliance-effect correlation, mean and variance of the compliance and effect) in infinite samples. However, as the situation under consideration gets more complicated, the bias formula quickly becomes intractable. The second part of the paper, therefore, uses simulations to demonstrate the relationship between the compliance-effect bias and various parameters, as well as the behavior of the estimated 2SLS standard errors. Furthermore, the simulation exercise assesses how the compliance-effect bias interacts with the finite sample bias when the analysis sample is small or when the instrument is weak. The paper also uses simulations to compare the properties of the 2SLS estimator with those of the ordinary least squares (OLS) estimator in the presence of the compliance-effect bias, the finite sample bias, and the omitted variable bias.

**Significance / Novelty of study:**

This is the first paper that systematically studies the form and behavior of the compliance-effect bias. It provides valuable insights to under what circumstances the compliance-effect correlation is likely to be problematic. It also compares the performances of the 2SLS estimator and the OLS estimator when various combinations of bias sources exist, thereby providing guidance to researchers as to which estimation method is more suitable for a given situation.

**Statistical, Measurement, or Econometric Model:**

The paper starts with the derivation of the formulas for the compliance-effect bias in the absence of the finite sample bias. In particular, it studies the following cases:

a. A single-site study with one mediator and one instrument
In this case, the following set of models are used to estimate the effect of the instrument, T, on the mediator, \( m^1 \), and the effect of T on the outcome, Y:

\[
\begin{align*}
    m^1_i &= y^1 T_i + \epsilon^1_i \\
    Y_i &= \beta^1 T_i + u_i 
\end{align*}
\]  

(1) \hspace{20ex} (2)

The model in Equation 1 is sometimes called the first-stage equation and the one in Equation 2 is referred to as the second-stage equation. It has been shown that the 2SLS estimator of the mediator effect on the outcome can be expressed as \( \hat{\delta}_1^{(2SLS)} = \frac{\hat{\beta}^1}{\gamma^1} \) (Wald estimator). In the absence of finite sample bias, it can be further shown that:

\[
E[\hat{\delta}_1^{(2SLS)}] = \delta^1 + \frac{\text{Cov}(y^1, \delta^1)}{\gamma^1} 
\]  

(3)

That is, the 2SLS estimator will be biased if the effect of the \( T \) on \( m^1 \) is correlated with the effect of \( m^1 \) on \( Y \) (the compliance-effect correlation). The compliance-effect correlation bias will be exacerbated when \( T \) is a weak instrument (or when \( y^1 \) is small).

b. A multiple-site study with one mediator

Since there are more sites than mediators in this case, there are at least three options to estimate the effect of mediator on the outcome.

\textbf{Option 1: Average of the within-site 2SLS estimates}

A separate 2SLS model like the one described in Equations 1 and 2 can be fitted within each site and the resulting within-site 2SLS estimates can be averaged across sites, weighting by sample sizes within sites. It can be shown that:

\[
E[\hat{\delta}_1^{(2SLS)}] = \delta^1 + \sum_s \frac{n_s}{N} \left( \frac{\text{Cov}_s(y^1, \delta^1)}{\gamma^1_s} \right) 
\]  

(4)

where \( n_s \) is the sample size for site \( s \), \( N \) is the total sample size across all sites, and \( \text{Cov}_s(y^1, \delta^1) \) is the compliance-effect correlation for site \( s \). Equation 4 shows that the average of the within-site 2SLS estimates will be a biased estimate of the average effect of \( M^1 \) on \( Y \) in the sampled population unless the second term above is zero.\(^6\)

\textbf{Option 2: 2SLS with site fixed effects and a single instrument}

Here the following model with a single instrument (the treatment indicator) and site fixed effects will be fitted to the pooled dataset:

\[
\begin{align*}
    m^1_{is} &= \Gamma^1_s + y^1 T_{is} + \epsilon^1_{is} \\
    Y_{is} &= B^1_s + \beta^1 T_{is} + u_{is} 
\end{align*}
\]  

(5) \hspace{20ex} (6)

\(^6\) The pooled 2SLS estimate is unbiased if the within-site compliance-effect correlation is either zero in all sites or positive in some sites and negative in others in such a way that the weighted average is zero, which is unlikely.
where $\Gamma^1_s$ and $B^1_s$ are the site fixed effects for the first and second-stage regression. As before, $\hat{\delta}^{1(2SLS)} = \frac{\hat{\beta}^1}{\hat{\gamma}^1}$. Assuming very large samples, it can be shown that:

$$E[\hat{\delta}^{1(2SLS)}] = \sum_s \frac{\gamma^1_s \delta^1_s}{\sum_s \gamma^1_s} \delta^1_s + \sum_s \frac{w_s \text{cov}_s(\gamma^1_t \delta^1_t)}{\sum_s w_s \gamma^1_s} \delta^1_s$$

(7)

where $w_s = \frac{n_s \sigma^2}{\sum_s n_s \sigma^2}$. That is, this approach yields a weighted average of the within-site average effects of $M^1$ on $Y$ and a bias-term, where the weights are proportional to the product of the sample size, the variance of the treatment, and the compliance within each site and the bias-term is a function of the within-site compliance-effect covariance.

If the variance of the treatment effect and the sample size is the same for all sites, (that is, if $w_s = 1$ for all $s$):

$$E[\hat{\delta}^{1(2SLS)}] = \sum_s \frac{\gamma^1_s \delta^1_s}{\sum_s \gamma^1_s} \delta^1_s + \frac{\text{cov}(\gamma^1_t, \delta^1_t)}{\gamma^1} + \sum_s \frac{w_s \text{cov}_s(\gamma^1_t \delta^1_t)}{\sum_s w_s \gamma^1_s} \delta^1_s$$

(8)

So the site fixed effects 2SLS estimator will be an unbiased estimate of the average effect of $M^1$ on $Y$ if the between-site compliance-effect correlation is zero and the average within-site compliance-effect correlation is zero. The extent of these biases will be exacerbated when the average compliance is low (weak instrument).

**Option 3: 2SLS with site fixed effects and site-by-treatment interactions as instruments**

A third option utilizes multiple site-by-treatment interactions to generate as many instruments as the number of sites ($S$):

$$m^1_{is} = \Gamma^1_s + \sum_{r=1}^S Y^1_r (I^r_s \cdot T_{is}) + \epsilon^1_{is}$$

$$Y_{is} = \Delta^1_s + \delta^1 m^1_{is} + u_{is}$$

(9)

(10)

where $I^r_s$ is a site indicator that equals one if individual $i$ is in site $r$ and zero otherwise. We then show that:

$$E[\hat{\delta}^{1(2SLS)}] = \sum_s \frac{w_s \gamma^1_s \delta^1_s}{\sum_s w_s \gamma^1_s} \left[ \delta^1_s + \frac{\text{cov}_s(\gamma^1_t \delta^1_t)}{\gamma^1_s} \right]$$

(11)

where $w_s$ is defined as in Equation 7. Equation 11 shows that this option yields a weighted average of the within-site average effects of $M^1$ on $Y$, where the weights are proportional to the product of the sample size, the variance of the treatment, and the square of the compliance within each site.

c. **A multiple-site study with multiple mediators**

Here we have multiple mediators and more sites than mediators (multiple site multiple instruments, or MSMM). Suppose that there are $S$ sites and $P$ mediators and that $S > P > 1$. This implies that there are $P$ first-stage and one second-stage equations:

$$m^p_{is} = \Gamma^p_s + \sum_{r=1}^S Y^p_r (I^r_s \cdot T_{is}) + \epsilon^p_{is} \quad (p=1,2,...,P)$$

(12)
\[ Y_{is} = \Delta_s + \sum_p \delta^p m^p_{is} + u_{is} \]  

(13)

It can be shown that in this case,

\[ \delta^p = \sum_s \left[ (w_s \gamma^p_s \sum_r \alpha_{pr} \gamma^r_s) \left( \delta^p_s + \frac{\text{cov}_{s}(y^p_s, \delta^p_s)}{\gamma^p_s} \right) \right] \]  

(14)

In other words, the MSMM 2SLS estimand is a weighted average of the site-specific \( \delta^p_s \)'s, where the weights include the sample size, the treatment variance, the site-specific compliance with mediator \( M^p \), and a weighted average of the site-specific compliances with each of the mediators. Unless the correlation of the site-specific effects with each of these weight terms is zero, the estimator will be biased. In addition, there is another source of bias in the MSMM 2SLS estimator – the ratio of the within-site mediator \( M^p \) compliance-effect covariance to the site-specific compliance (which is also weighted by the same factors as above).

These analyses show that the 2SLS may yield biased estimates of the effects of mediator(s) on the outcome. There are several sources of bias: within-site compliance-effect correlation (i.e., individuals whose mediator values are most strongly affected by the treatment (instrument) respond more, on average, to the mediators); between-site compliance-effect correlations; and unequal treatment variance across sites.

**Usefulness / Applicability of Method:**

The paper utilizes simulated data corresponding to the situations described above to facilitate the understanding of the derived expressions for the compliance-effect bias in the 2SLS estimator. It further demonstrates how substantial the compliance-effect bias can be in different situations and the relationship between the bias and the various parameters that affect the bias. For example, it shows that for the multiple-sites two mediators case, for certain parameter values, holding other parameters fixed, the 2SLS bias for the first mediator tends to increase as:

- the strength of the instrument decreases,
- the variance of the effect of treatment on mediator increases,
- the effect of mediator on outcome increases, or
- the compliance-effect correlation (either within a mediator or cross-mediators) increases.

The paper also relies on simulations to study the standard error of the 2SLS estimator in the presence of just compliance-effect bias, just finite sample bias, or both. The simulated 2SLS estimates are also compared to the corresponding OLS estimates to assess which estimation method produces less bias or is more efficient under given conditions.

**Conclusions:**

This paper derives the expressions for the bias in the 2SLS estimator when the effects of treatment on mediator(s) are correlated with the effects of mediator(s) on the outcome in various situations and uses simulated data to demonstrate the behavior of the compliance-effect bias. It shows that the compliance-effect bias can be substantial under certain conditions. Therefore it is important for researchers to assess the potential magnitude of this bias before selecting the method to conduct mediational analyses.
Appendices

Not included in page count.

Appendix A. References
References are to be in APA version 6 format.


Title: Mediation and Spillover Effects in Group-Randomized Trials with Application to the 4Rs Evaluation

Author(s):
Tyler J. VanderWeele (Harvard School of Public Health)  
tvanderw@hsph.harvard.edu
Guanglei Hong (University of Chicago, contact author)  
ghong@uchicago.edu
Stephanie M. Jones (Harvard Graduate School of Education)  
jonesst@gse.harvard.edu
Joshua L. Brown (Fordham University)  
joshua.brown@nyu.edu
Abstract Body
Limit 5 pages single spaced.

Background / Context:
Description of prior research and its intellectual context.

A common setting in educational research consists of a randomized intervention at the school level, a mediator of interest at the classroom or teacher level and an outcome of interest at the child level. A common approach to addressing mediation in such settings consists of regressing the outcome on the treatment with and without the mediator variable. This approach to mediation analysis is subject to several limitations. First, the approach ignores selection into mediator levels; although the treatment is randomized, the mediator is not and thus analyses ignoring this selection issue are subject to potentially severe biases due to confounding (Judd and Kenny, 1981; Robins and Greenland, 1992; Pearl, 2001). The second issue with the standard regression approach is that potential interaction between the effects of treatment and the mediator on the outcome are typically ignored. Recent literature on causal inference has made clear that mediation analysis becomes considerably more complex when such interactions are present (Pearl, 2001). A third issue with the standard regression approach is that it ignores issues of interference and spillover effects. Child level outcomes may depend not only on the characteristics of the child’s own classroom but also on the characteristics of other classrooms because of social interactions among children from different classrooms. The issue is referred to as one of interference between units in the statistics literature (Cox, 1958). No interference between units is a component of Rubin’s Stable Unit Treatment Value Assumption or SUTVA (Rubin, 1980, 1986). The assumption will be violated in settings in which social interactions allow one individual’s exposure to affect the outcomes of other individuals. Such interference is part of the theoretical rationale of the 4Rs program which focuses on bringing together educators’ collective efforts within a school. Analyses of causal effects are considerably more complex in the face of such interference.

Purpose / Objective / Research Question / Focus of Study:
Description of the focus of the research.

In this paper we extend recent work on mediation in a multilevel setting (VanderWeele, 2010) and on causal inference under interference among units (Hong and Raudenbush, 2006; Hudgens and Halloran, 2008; Rosenbaum, 2007; Sobel, 2006) to develop a template for the mediation analysis of group randomized educational interventions. The present work will contribute to the literature on interference, in particular on interference in the context of mediation analysis. We will show that not only does the total effect of the intervention decompose into a direct effect and an indirect effect mediated through classroom quality but also that the indirect effect itself decomposes into an effect mediated through the quality of a child’s own classroom and a spillover effect from the quality of the other classrooms at a school. We will give some results for the identification of these direct, indirect and spillover effects and consider the consequences of ignoring interference when it is in fact present. We will then analyze the effects of the Reading, Writing, Respect and Resolution (4Rs) intervention in a group randomized trial.

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.
Peer influence and social interactions can give rise to spillover effects in which characteristics of one individual unit may affect outcomes of other individual units. Evaluators who choose groups rather than individuals as experimental units in group randomized trials often anticipate that the desirable changes in targeted social behaviors will be reinforced through interference among individuals in a group exposed to the same treatment. Failure to account for such spillover effects can result in bias and problems with interpretation. Using a counterfactual conceptualization of direct, indirect and spillover effects, we provide a framework that can accommodate issues of mediation and spillover effects in group randomized trials.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

In this paper, we are interested in the extent to which the effect of the 4Rs intervention on child outcomes is mediated by classroom quality. Let \( T_k \) denote the school-wide randomized treatment for school \( k \) (1 for the 4Rs intervention; 0 for control). Let \( M_{jk} \) denote the classroom level mediator for classroom \( j \) in school \( k \). In the 4Rs intervention study the mediator of interest is a continuous measure of classroom quality. Let \( J_k \) denote the number of classrooms in school \( k \). Let \( Y_{ijk} \) denote the child-level outcome for child \( i \) in classroom \( j \) and school \( k \). In the 4Rs study this outcome is a continuous score measuring depressive symptoms. Let \( Y_{ijk}(t_k) \) denote the potential or counterfactual outcome that child \( i \) in classroom \( j \) and school \( k \) would have obtained if the school-level treatment, \( T_k \), were set to \( t_k \). Similarly, let \( M_{jk}(t_k) \) denote the potential or counterfactual mediator that classroom \( j \) in school \( k \) would have obtained if the school-level treatment, \( T_k \), were set to \( t_k \). We assume that children do not change schools as a result of the treatment to which a particular school is assigned. Hong and Raudenbush (2006) referred to this assumption as that of "intact clusters." We also assume that there is no interference between schools (i.e. that the treatment received at one school does not affect the outcomes of the children at any other schools). To incorporate within-school interference into our potential outcomes notation, we let \( Y_{ijk}(t_k, m_{jk}, A_{jk}) \) denote the counterfactual outcome that child \( i \) in classroom \( j \) and school \( k \) would have obtained if the school-level treatment in school \( k \) were set to \( t_k \), if the quality in classroom \( j \) of school \( k \) were set to \( m_{jk} \) and if the quality of all other classrooms in school \( k \) were set to the vector \( A_{jk} = (m_{1k}, \ldots, m_{j-1k}, m_{j+1k}, \ldots, m_{Jk}) \). Following Hong and Raudenbush (2006) and Hudgens and Halloran (2008), we assume that the potential outcome \( Y_{ijk}(t_k, m_{jk}, A_{jk}) \) depends on \( A_{jk} \) through some scalar function \( G(A_{jk}) \) of \( A_{jk} \) so that we may express the potential outcome as \( Y_{ijk}(t_k, m_{jk}, G(A_{jk})) \). For example, \( G(A_{jk}) \) may denote the average quality for all classrooms in school \( k \) other than classroom \( j \). Here we let \( Y_{ijk}(t, m, g) \) denote the outcome for child \( i \) in classroom \( j \) and school \( k \) if the school received treatment \( t \), the child’s classroom had quality \( m \), and the scalar function of the quality of other classrooms, \( G(A_{jk}) \), took the value \( g \).

The causal contrast \( E[Y_{ijk}(1, m, g) - Y_{ijk}(0, m, g)] \) captures the direct effect of the 4Rs program but also intervening to fix the quality of the child’s own classroom to level \( m \) and intervening to fix the average quality of other classrooms to \( g \). This quantity is referred to as a controlled direct effect of treatment. Likewise the contrast \( E[Y_{ijk}(t, m, g) - Y_{ijk}(t, m^*, g)] \) could be used to assess the effect of a child’s own classroom quality (comparing levels \( m \) and \( m^* \)) on a child’s outcome and to examine whether the contrast varies with \( t \) or \( g \). Similarly, the contrast \( E[Y_{ijk}(t, m, g) - Y_{ijk}(t, m, g^*)] \) could be used to assess the spillover effect of the quality of classrooms other than the child’s own and whether the contrast varies with \( t \) or \( m \).
When the classroom quality mediators are set to the levels they would have been at under the control condition, the natural direct effect is defined as $E[Y_{ijk}(1, M_{jk}(0), G(M_{jk}(0))) - Y_{ijk}(0, M_{jk}(0), G(M_{jk}(0)))].$ We can also define a natural indirect effect as $E[Y_{ijk}(1, M_{jk}(1), G(M_{jk}(1))) - Y_{ijk}(1, M_{jk}(0), G(M_{jk}(0)))].$ As in the case of non-clustered treatments without interference (Pearl, 2001), the total effect of the intervention on the outcome $E[Y_{ijk}(1) - Y_{ijk}(0)]$ decomposes into natural direct and indirect effects. The decomposition will hold even if there are interactions between the effects of the treatment and the mediator on the outcome. The natural indirect effect further decomposes into a within-classroom mediated effect $E[Y_{ijk}(1, M_{jk}(1), G(M_{jk}(0))) - Y_{ijk}(1, M_{jk}(0), G(M_{jk}(0)))$ and a spillover mediated effect $E[Y_{ijk}(1, M_{jk}(1), G(M_{jk}(1))) - Y_{ijk}(1, M_{jk}(1), G(M_{jk}(0))))].$

We will let $X_{ijk}$, $W_{jk}$, and $V_k$ denote child-level, class-level, and school-level baseline covariates, respectively. We will use $X_{ijk}$ to denote the vector of child-level baseline covariates for children in school $k$ other than child $i$ in classroom $j$, $W_{jk}$ to denote classroom-level baseline covariates for classrooms in school $k$ other than classroom $j$. We will consider certain functions of the baseline covariates of other children in the classroom (or even at the school), $h_1(X_{ijk})$, and of baseline covariates of classrooms other than a child’s own $h_2(W_{jk}).$ To simplify notation we let $L_{ijk} = (X_{ijk}, W_{jk}, V_k, h_1(X_{ijk}), h_2(W_{jk})).$

**Usefulness / Applicability of Method:**

*Demonstration of the usefulness of the proposed methods using hypothetical or real data.*

We present four identification results, one for controlled direct effects, one for natural direct and indirect effects, one for the spillover and within-classroom mediated effects, and one for the consequences of ignoring interference when it is in fact present. For sets of random variables $A$, $B$, and $C$, we will use $A \perp B \mid C$ to represent that $A$ is independent of $B$ conditional on $C$.

**Theorem 1.** If for all $t$, $m$, $g$ we have that $Y_{ijk}(t, m, g) \perp T_k \mid L_{ijk}$ and that $Y_{ijk}(t, m, g) \perp \{M_{jk}, G(M_{jk})\} \mid T_k, L_{ijk}$, then we can identify the controlled direct effect of the treatment and that of each mediator.

**Theorem 2.** If in addition to the assumptions stated in Theorem 1, we also have that $\{M_{jk}(t), G(M_{jk}(t))\} \perp T_k \mid L_{ijk}$ and that for all $t, t^*, m, g, Y_{ijk}(t, m, g) \perp \{M_{jk}(t^*), G(M_{jk}(t^*))\} \mid L_{ijk}$, then we can identify the natural direct effect and the natural indirect effect.

**Theorem 3.** If in addition to the assumptions stated in Theorems 1 and 2, we also have that for $t' \neq t^*$, $M_{jk}(t') \perp G(M_{jk}(t^*)) \mid L_{ijk}$, then we can identify the within-class mediated effect and the spillover mediated effect.

**Theorem 4.** Suppose that the assumptions stated in Theorems 1, 2, and 3 hold. And suppose we also require that for all $t$, $M_{jk}(t) \perp G(M_{jk}(t)) \mid L_{ijk}$, then we can ignore interference while still obtaining an estimate of the within-classroom mediated effect and obtaining the sum of a spillover mediated effect and the actual natural direct effect. However, even if all the above assumptions hold, if the substantive question of interest is whether classroom quality mediates the effect of treatment, ignoring interference would lead to an underestimate of the actual
importance of classroom quality since it will not include an assessment of the effect mediated through the quality of other classrooms.

**Research Design:**
*Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).*
*(May not be applicable for Methods submissions)*

The 4R[s program is a school-based intervention in literacy development, conflict resolution, and intergroup understanding. The study (Jones, Brown, & Aber, in press) involved a 3-year, 6-wave longitudinal experimental design with measurements in the fall and spring semester of each year. The eighteen New York City elementary schools in the study were fairly representative of the demographic characteristics of New York City schools and included 923 students in 82 classrooms. The schools were pair matched based on twenty school characteristics including size, reading achievement, race/ethnic composition, mobility/two-year stability, school lunch receipt, expenditures, attendance and organizational readiness. Within each pair, schools were randomly assigned to either the 4Rs treatment or the control group. The intervention was implemented school-wide from grades K-6 for 3 years. All 3rd grade children in each school were followed over three years through 5th grade. In the application here, we will consider the first year of the study for the children beginning in third grade.

**Data Collection and Analysis:**
*Description of the methods for collecting and analyzing data.*
*(May not be applicable for Methods submissions)*

Classroom quality was measured in the spring semester using the CLASS scoring system (Pianta, La Paro, and Hamre, 2005) which assesses instructional support, emotional support, and organizational climate with an overall score between 1 and 7. We dichotomized this measure using 4.4 as the cutoff. The child-level outcome was depressive symptoms scored on a scale of 0 to 1. Covariates in the model were chosen based on prior empirical work (Brown, Jones, LaRusso, & Aber, 2010). The covariates were at least marginally predictive of either the outcome or the mediator. The models also included pair fixed effects to control for school-level factors. We fitted a multilevel model for the effect of treatment on the depressive symptoms, a multilevel model for the effect of treatment on class quality, and finally a multilevel model for the effects of treatment, classroom quality, and quality of other classrooms on depressive symptoms with the interactions between these variables saturated. The parameter values and model-based standard errors were estimated via maximum likelihood in HLM 6.0.

**Findings / Results:**
*Description of the main findings with specific details.*
*(May not be applicable for Methods submissions)*

The estimated treatment effect of the 4Rs intervention on depressive symptoms is -0.052 (s.e. = 0.023, $t = -2.29, p = 0.05$), suggesting a marginally significant effect of the treatment in reducing child depressive symptoms. The estimated treatment effect of the 4Rs intervention on classroom quality is 0.45 (s.e. = 0.20, $t = 2.28, p = 0.05$). In the control schools it appears that depressive symptoms are highest for children in classrooms in which the quality of the child’s own classroom is low but the quality of other classrooms at the school is relatively high. Apparently it
is also only in this type of classrooms that the 4Rs intervention has a statistically significant direct effect on depressive symptoms (see Table 1). However, a $\chi^2$ test failed to reject the null hypothesis of no interaction of treatment with either a child’s own classroom quality or the quality of other classrooms. We then obtained an estimate of the overall controlled direct effect of treatment of -0.058 ($p = 0.025$). If we take the $\chi^2$ test result as an indication that the assumptions in Theorem 2 holds, then this estimate would also correspond to the natural direct effect. We could obtain a natural indirect effect by subtracting the natural direct effect from the total effect of treatment which would give $-0.052 - (-0.058) = 0.006$. This provides very little evidence that any of the effect of the 4Rs intervention on depressive symptoms is mediated through either the quality of a child’s own classroom or the quality of the classes other than the child’s own. Because this particular evaluation of the 4Rs intervention was powered to be able to assess only the total effect, not direct and indirect effects, our results here are at best suggestive.

Conclusions:
Description of conclusions, recommendations, and limitations based on findings.

In this paper we have made a number of contributions to allow researchers to address questions of mediation and spillover effects in group randomized trials. The approach we have described here constitutes an advance over the standard approach to estimating direct and indirect effects that is often used in group-randomized trials. Specifically, our approach (i) makes explicit the assumptions required for identification that will be important in study design and data analysis of group randomized trials, (ii) accommodates possible interactions that may be present, (iii) allows for interference between individuals in different clusters (e.g., classrooms in the 4Rs evaluation), and (iv) allows for the definition, identification, and estimation of spillover effects. In particular, by relaxing the no-interference assumption, we have been able to investigate spillover effects that will often be of substantive and theoretical interests. Interference is not simply a problem that must be dealt with but in fact gives rise to research questions about spillover effects that are of interest in their own right. In addition, we have provided an analysis of the mediation and spillover effects in the 4Rs evaluation. The chief limitations of the analysis are: (i) a relatively large sample size may be required to draw reliable inferences about mediation and spillover effects; and (ii) relatively strong identification assumptions are required to empirically estimate these effects from data.

The approach that we have presented here could be extended in a number of directions. First, future work could consider accommodating longitudinal settings as the mediator and outcome changes over time. Second, work has been done on using weighting techniques (van der Laan and Petersen, 2008; VanderWeele, 2009; Hong, 2010) rather than regression to address confounding control in questions of mediation analysis; future research could attempt to extend these weighting techniques to estimate and distinguish spillover mediated effects and within-classroom mediated effects. Third, further research could develop sensitivity analysis techniques to assess the extent to which an unobserved variable affecting both the mediator and the outcome (and thus giving rise to confounding of the effects of both the mediator in a child’s own classroom and that of the mediator in the other classrooms) might invalidate the inference about direct, indirect and spillover effects.
Appendices
Not included in page count.

Appendix A. References
References are to be in APA version 6 format.


Appendix B. Tables and Figures

Not included in page count.

Table 1
Controlled Direct Effects of 4Rs Program by Class Quality Indicators

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>M=0, G=0</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.58</td>
<td>0.03</td>
<td>12.28***</td>
</tr>
<tr>
<td>Direct effect</td>
<td>-0.05</td>
<td>0.05</td>
<td>-1.18</td>
</tr>
<tr>
<td><strong>M=0, G=1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.64</td>
<td>0.06</td>
<td>10.16***</td>
</tr>
<tr>
<td>Direct effect</td>
<td>-0.13</td>
<td>0.05</td>
<td>-2.63*</td>
</tr>
<tr>
<td><strong>M=1, G=0</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.58</td>
<td>0.06</td>
<td>10.08***</td>
</tr>
<tr>
<td>Direct effect</td>
<td>-0.01</td>
<td>0.05</td>
<td>-0.14</td>
</tr>
<tr>
<td><strong>M=1, G=1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.59</td>
<td>0.09</td>
<td>6.36***</td>
</tr>
<tr>
<td>Direct effect</td>
<td>-0.04</td>
<td>0.07</td>
<td>-0.56</td>
</tr>
</tbody>
</table>

*p < .05; ***p < .001